Measuring Data Leakage in Machine-Learning Models with Fisher Information (Extended Abstract)*

Awni Hannun[1†], **Chuan Guo**[2], **Laurens van der Maaten**[2]
[1]Zoom AI
[2]Meta AI Research
awni.hannun@zoom.us, {chuanguo, lvdmaaten}@fb.com

Abstract

Machine-learning models contain information about the data they were trained on. This information leaks either through the model itself or through predictions made by the model. Consequently, when the training data contains sensitive attributes, assessing the amount of information leakage is paramount. We propose a method to quantify this leakage using the Fisher information of the model about the data. Unlike the worst-case *a priori* guarantees of differential privacy, *Fisher information loss* measures leakage with respect to specific examples, attributes, or sub-populations within the dataset. We motivate Fisher information loss through the Cramér-Rao bound and delineate the implied threat model. We provide efficient methods to compute Fisher information loss for output-perturbed generalized linear models. Finally, we empirically validate Fisher information loss as a useful measure of information leakage.

1 Introduction

Machine-learning models trained on sensitive data are often made public. Even when the models are not explicitly released, they may be implicitly leaked from their predictions [Papernot *et al.*, 2017; Tramèr *et al.*, 2016]. Undeniably, these models contain information about the data they were trained on. Without mitigating measures, training set membership can often be inferred [Shokri *et al.*, 2017; Yeom *et al.*, 2018], and sensitive attributes or even whole examples can be extracted [Carlini *et al.*, 2019; Fredrikson *et al.*, 2014].

Assessing the information leaked from models about their training data is commonly done with techniques such as differential privacy [Dwork *et al.*, 2006]. Differential privacy successfully avoids the "just a few" failure mode of more heuristic privacy assessments, in which privacy is protected for many but not all individuals [Dwork and Roth, 2014]. This comes at the cost of a worst-case assessment, leading to large differences in the vulnerability of individuals to privacy attacks. Also, a mismatch exists between the protection of differential privacy, which is relative to participation in a dataset, and privacy attacks, which can take advantage of the absolute information leaked from a trained model [Carlini *et al.*, 2019; Long *et al.*, 2018]. Furthermore, differential privacy degrades when correlations exist in the dataset [Ghosh and Kleinberg, 2016; Liu *et al.*, 2016].

We propose an example-specific and correlation-aware measure of data leakage using Fisher information. The quantity, which we term *Fisher information loss*, can assess the leakage of a model about various subsets of the full dataset. This includes, for example, assessments at the granularity of individual attributes, individual examples, groups of examples, or the full training set. We show, via the Cramér-Rao bound, that under specific assumptions the ability of an adversary to estimate the underlying data from a model with bounded Fisher information loss is limited. We also demonstrate that, unlike differential privacy, Fisher information loss does not implicitly degrade when data is correlated.

The ability of Fisher information loss to measure per-example privacy loss means it can be used as the basis for algorithms that aim to achieve fairness in privacy [Cummings *et al.*, 2019; Ekstrand *et al.*, 2018]. We demonstrate this by developing an algorithm that balances Fisher information loss for individuals in the training set, thereby resolving the problem that subgroups may have "disparate vulnerability" to privacy attacks [Yaghini *et al.*, 2019].

2 Related Work

This work builds on and complements a significant body of prior work in assessing the privacy of a model with respect to the data it was trained on. This includes approaches which obfuscate the original data such as k-anonymity [Samarati and Sweeney, 1998], information theoretic criteria [Agrawal and Aggarwal, 2001], and, perhaps the most commonly studied, differential privacy [Dwork *et al.*, 2006] and its more recent variations [Mironov, 2017; Dong *et al.*, 2019]. While these methods often provide rigorous privacy guarantees, most of them do not explicitly bound the *inferential power* of an adversary, and correlations in the training dataset can be exploited by an adversary to mount a successful inference attack [Machanavajjhala *et al.*, 2007; Li *et al.*, 2007; Liu *et al.*, 2016]. Moreover, these definitions do not offer

[*] Best Paper at the 37th Conference on Uncertainty in Artificial Intelligence (2021).

[†] Contact Author

simple and flexible approaches to measure the information leaked by a model about varying subsets of the training data. This can lead to unequal vulnerability to privacy attacks despite attempts to protect privacy [Yaghini *et al.*, 2019].

Because these privacy assessment techniques do not identify vulnerability at the level of the individual or subpopulation, prior work exists to quantify the susceptibility of such subgroups to privacy attacks. Farokhi and Kaafar [2020] propose an information theoretic measure of the vulnerability of individual examples to membership inference attacks. Carlini *et al.* [2019] propose a heuristic which can infer the susceptibility of data to model inversion attacks.

This work also builds upon prior studies of Fisher information as a measure of privacy. An early study analyzed loss of Fisher information in a general randomized response framework [Anderson, 1977]. More recently, the relationship of privacy with Fisher information to estimator error through the use of the Cramér-Rao bound was investigated [Farokhi and Sandberg, 2017]. In particular, Farokhi and Sandberg [2017] turn to Fisher information as a practical alternative to differential privacy in protecting data collected from smart power meters. We build from these works in several directions, including a broader application to generalized linear models as well as the use of Fisher information loss in an algorithm which can provide fairness in data leakage.

3 Fisher Information Loss

Let $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$ be a training dataset of n examples with $\boldsymbol{x}_i \in \mathbb{R}^d$ and scalar target y_i. We denote by $\mathcal{A}(\mathcal{D})$ a randomized learning algorithm which outputs a hypothesis h from a predefined hypothesis space $\mathcal{H}$. Treating the hypothesis as a random variable, $p_{\mathcal{A}}(h \mid \mathcal{D})$ is the probability density of h given $\mathcal{D}$ for the randomized algorithm $\mathcal{A}(\mathcal{D})$. We denote by $\mathcal{I}_h(\mathcal{D}) \in \mathbb{R}^{n(d+1)\times n(d+1)}$ the Fisher information matrix (FIM) defined by

$$\mathcal{I}_h(\mathcal{D}) = -\mathbb{E}_h \left[\nabla^2_{\mathcal{D}} \log p_{\mathcal{A}}(h \mid \mathcal{D})\right] \tag{1}$$

where $\nabla^2_{\mathcal{D}}$ yields the matrix of second derivatives of $\log p_{\mathcal{A}}(h \mid \mathcal{D})$ with respect to the values in $\mathcal{D}$, and the expectation is taken over the randomness in $\mathcal{A}(\mathcal{D})$. The FIM, $\mathcal{I}_h(\mathcal{D})$, is a measure of the information that the hypothesis h contains about the training data $\mathcal{D}$. Hence, we use the FIM to measure the information loss from releasing the output h of a single evaluation of $\mathcal{A}(\mathcal{D})$.

Definition (Fisher information loss)**.** *We say that $h \sim \mathcal{A}(\mathcal{D})$ has Fisher information loss (FIL) of η with respect to $\mathcal{D}$ if*

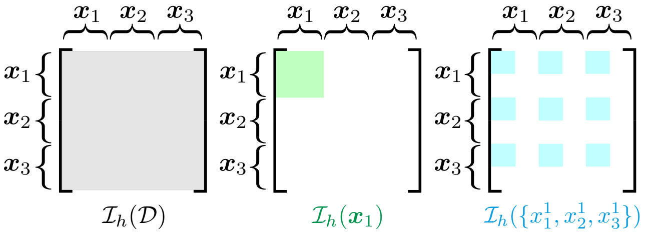
$$\|\mathcal{I}_h(\mathcal{D})\|_2 \leq \eta^2, \tag{2}$$

where $\|\mathcal{I}_h(\mathcal{D})\|_2$ denotes the 2-norm, or largest singular value, of the FIM. A smaller η means h contains less Fisher information about the training data, $\mathcal{D}$.

Motivation. Fisher information is a classical tool used in statistics for lower bounding the variance of an estimator. We utilize this property to demonstrate that a small FIL implies a large variance for any unbiased estimate of the data.

Let $\boldsymbol{z} = [\boldsymbol{x}_1^\top, y_1, \ldots, \boldsymbol{x}_n^\top, y_n] \in \mathbb{R}^{n(d+1)}$ be the vector formed by concatenating the examples in $\mathcal{D}$, and $z \in \boldsymbol{z}$ an arbitrary element. If the FIL of h with respect to $\mathcal{D}$ is bounded by η, then for any unbiased estimator $\hat{z}$ of z, we have:

$$\mathrm{Var}(\hat{z}) \geq \frac{1}{\eta^2}. \tag{3}$$


$\boldsymbol{x}_1$ $\boldsymbol{x}_2$ $\boldsymbol{x}_3$
$\boldsymbol{x}_1$
$\boldsymbol{x}_2$
$\boldsymbol{x}_3$
$\mathcal{I}_h(\mathcal{D})$
$\mathcal{I}_h(\boldsymbol{x}_1)$
$\mathcal{I}_h(\{x_1^1, x_2^1, x_3^1\})$

(a) Full FIM (b) Example-level FIM (c) Attribute-level FIM

Figure 1: The FIM for different subsets of the training data.

Hence, a smaller FIL implies a larger variance in any unbiased attempt to infer z. Equation 3 follows directly from the Cramér-Rao bound. Indeed, under a fairly relaxed regularity condition [Kay, 1993], for any unbiased estimator $\hat{\boldsymbol{z}}$ of the data $\boldsymbol{z}$, the Cramér-Rao bound states that:

$$\mathbb{E}\left[(\hat{\boldsymbol{z}} - \boldsymbol{z})(\hat{\boldsymbol{z}} - \boldsymbol{z})^\top\right] \succeq \mathcal{I}_h(\boldsymbol{z})^{-1}, \tag{4}$$

where $A \succeq B$ if the matrix $A - B$ is positive semidefinite. Since the estimator is unbiased, this implies the covariance of $\hat{\boldsymbol{z}}$ is similarly bounded:

$$\mathrm{Cov}(\hat{\boldsymbol{z}}) \succeq \mathcal{I}_h(\boldsymbol{z})^{-1}. \tag{5}$$

If $\mathcal{A}(\mathcal{D})$ has an FIL of η then equation 2 implies $\mathcal{I}_h(\boldsymbol{z})^{-1}_{ii} \geq 1/\eta^2$ for all i, and from equation 5, $\mathrm{Var}(\hat{z}) \geq 1/\eta^2$ follows.

Correlated data. Fisher information loss also provides some security in the presence of intra-dataset correlations. If a model has an FIL bounded by η with respect to the training data (in vector form $\boldsymbol{z}$), then the covariance matrix for any unbiased estimator $\hat{\boldsymbol{z}}$ of that data is bounded by $\|\mathrm{Cov}(\hat{\boldsymbol{z}})\|_2 \geq 1/\eta^2$. This limits the ability of an adversary using an unbiased estimator to infer relative differences between elements in the dataset.

3.1 Properties of FIL

Subsets. In many cases we are interested in measuring FIL with respect to a single example $(\boldsymbol{x}_i, y_i) \in \mathcal{D}$. In this case, we can compute the example-specific FIM by selecting the corresponding entries of the full FIM. Given the FIM for the vector represented data $\boldsymbol{z}$, computing the FIM for the i-th example amounts to selecting the submatrix of size $(d+1) \times (d+1)$ with upper left corner $\mathcal{I}_h(\boldsymbol{z})_{i(d+1),i(d+1)}$. Similarly, computing the FIM for a specific attribute over all examples, $\{x_i^j\}_{i=1}^n$, amounts to constructing the submatrix by selecting the corresponding entries from the full FIM. See figure 1 for an illustration.

Composition. The Fisher information, and hence FIL, compose additively. Given k independent evaluations of $\mathcal{A}(\mathcal{D})$ each with an FIL of η, the combined FIL is at most $\sqrt{k}\eta$. More generally, given an evaluation of k unique but independent randomized algorithms $\{h_i \sim \mathcal{A}_i(\mathcal{D}) \mid i = 1, \ldots, k\}$ each with an FIL of η_i, the combined FIL for all h_i about $\mathcal{D}$ is at most $(\sum_{i=1}^k \eta_i^2)^{1/2}$.

Algorithm 1 Iteratively reweighted Fisher information loss.

```
1: Input: Data set $\mathcal{D}$, loss function $\ell(\cdot)$, number of iterations $T$, and noise scale $\sigma$.
2: Initialize sample weights $\omega_i^0 \leftarrow 1$.
3: for $t \leftarrow 1$ to $T$ do
4:     $\boldsymbol{w}^* \leftarrow \arg\min_{\boldsymbol{w}} \sum_{i=1}^n \omega_i^{t-1} \ell(\boldsymbol{w}^\top \boldsymbol{x}_i, y_i) + \frac{n\lambda}{2}\|\boldsymbol{w}\|_2^2$.
5:     $\boldsymbol{w}' \leftarrow \boldsymbol{w}^* + \boldsymbol{b}$ where $\boldsymbol{b} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I})$.
6:     $\eta_i \leftarrow (\|\mathcal{I}_{\boldsymbol{w}'}(\boldsymbol{x}_i, y_i)\|_2)^{1/2}$.
7:     $\omega_i^t \leftarrow \frac{n\omega_i^{t-1}/\eta_i}{\sum_{i=1}^n \omega_i^{t-1}/\eta_i}$.
8: end for
9: Return: The private weights $\boldsymbol{w}'$.
```

Closed under post-processing. As in differential privacy, FIL is closed under post-processing. If $\|\mathcal{I}_h(\mathcal{D})\|_2 \leq \eta^2$, then for any function of the hypothesis, $g(h)$, $\|\mathcal{I}_{g(h)}(\mathcal{D})\|_2 \leq \eta^2$.

3.2 Computing FIL

Learning setting. We assume a linear model with parameters $\boldsymbol{w} \in \mathbb{R}^d$ and minimize the regularized empirical risk:

$$\boldsymbol{w}^* = f(\mathcal{D}) \overset{\text{def}}{=} \arg\min_{\boldsymbol{w}} \sum_{i=1}^n \ell(\boldsymbol{w}^\top \boldsymbol{x}_i, y_i) + \frac{n\lambda}{2}\|\boldsymbol{w}\|_2^2. \quad (6)$$

Furthermore, we assume that the loss $\ell(\boldsymbol{w}^\top \boldsymbol{x}_i, y_i)$ is convex and twice differentiable. We denote by $f(\mathcal{D})$ the minimizer, $\boldsymbol{w}^*$, of equation 6 as a function of the dataset $\mathcal{D}$. When computing FIL at the example level, we let $f_i(\boldsymbol{x}, y)$ be the minimizer of equation 6 as a function of the i-th data point:

$$f_i(\boldsymbol{x}, y) \overset{\text{def}}{=} \arg\min_{\boldsymbol{w}} \sum_{j \neq i} \ell(\boldsymbol{w}^\top \boldsymbol{x}_j, y_j) + \ell(\boldsymbol{w}^\top \boldsymbol{x}, y) + \frac{n\lambda}{2}\|\boldsymbol{w}\|_2^2. \quad (7)$$

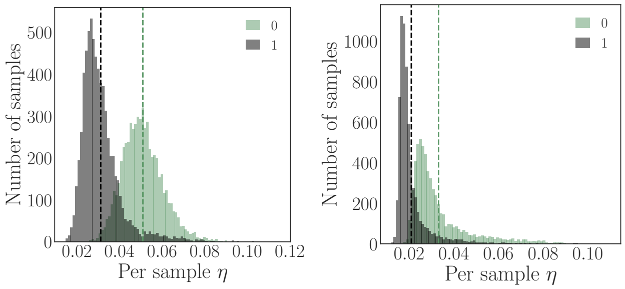
Output perturbation. The definition of FIL in equation 3 only applies to a randomized learning algorithm $\mathcal{A}$ with a differentiable density function. To obtain such a randomized learning algorithm from the minimizer $\boldsymbol{w}^*$ in equation 6, we adopt the Gaussian mechanism from differential privacy [Dwork *et al.*, 2006]. The Gaussian mechanism adds zero-mean isotropic Gaussian noise to the parameters $\boldsymbol{w}^*$. Let $\mathcal{A}(\mathcal{D}) = f(\mathcal{D}) + \boldsymbol{b}$ be the output-perturbed learning algorithm, where $\boldsymbol{b} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I})$. The FIM of $\boldsymbol{w}' \sim \mathcal{A}(\mathcal{D})$ is given by:

$$\mathcal{I}_{\boldsymbol{w}'}(\mathcal{D}) = \frac{1}{\sigma^2} \boldsymbol{J}_f^\top \boldsymbol{J}_f, \quad (8)$$

where $\boldsymbol{J}_f \in \mathbb{R}^{d \times n(d+1)}$ is the Jacobian of $f(\mathcal{D})$ with respect to $\mathcal{D}$. The Jacobian $\boldsymbol{J}_f$ captures the sensitivity of the minimizer $\boldsymbol{w}^*$ with respect to the training dataset $\mathcal{D}$. The FIL of the Gaussian mechanism with scale σ is then:

$$\eta = \frac{1}{\sigma}\|\boldsymbol{J}_f\|_2. \quad (9)$$

Jacobian of the minimizer. We aim to compute the Fisher information loss of the Gaussian mechanism in equation 9 at the example level. This requires the Jacobian $\boldsymbol{J}_{f_i}$ of $f_i(\boldsymbol{x}, y)$, equation 7, with respect to $(\boldsymbol{x}, y)$ evaluated at $(\boldsymbol{x}_i, y_i)$. For a convex, twice-differentiable loss function $\ell(\boldsymbol{w}^\top \boldsymbol{x}, y)$, the Jacobian $\boldsymbol{J}_{f_i}$ evaluated at $(\boldsymbol{x}_i, y_i)$ is given by:



(a) Linear

(b) Logistic

Figure 2: Histograms of per-example η separated by class label for the MNIST training set for both linear and logistic regression. Each class label's mean η is denoted by the dashed vertical line.

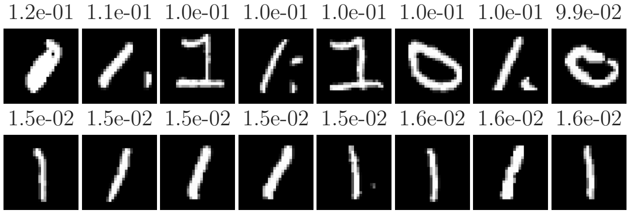


Figure 3: The eight images with the smallest and largest η over the MNIST training set using linear regression. The number above each individual image is the corresponding η.

$$\boldsymbol{J}_{f_i}\Big|_{\boldsymbol{x}_i, y_i} = -\boldsymbol{H}_{\boldsymbol{w}^*}^{-1} \nabla_{\boldsymbol{x}, y} \nabla_{\boldsymbol{w}} \ell(\boldsymbol{w}^{*\top} \boldsymbol{x}_i, y_i). \quad (10)$$

The Hessian is computed over the full dataset $\boldsymbol{H}_{\boldsymbol{w}^*} = \sum_{i=1}^n \nabla_{\boldsymbol{w}}^2 \ell(\boldsymbol{w}^{*\top} \boldsymbol{x}_i, y_i) + n\lambda \boldsymbol{I}$. The term $\nabla_{\boldsymbol{x}, y} \nabla_{\boldsymbol{w}} \ell = [\nabla_{\boldsymbol{x}} \nabla_{\boldsymbol{w}} \ell, \nabla_y \nabla_{\boldsymbol{w}} \ell] \in \mathbb{R}^{d \times (d+1)}$ is the Jacobian of $\nabla_{\boldsymbol{w}} \ell$ with respect to $(\boldsymbol{x}, y)$ with entries given by:

$$(\nabla_{\boldsymbol{x}} \nabla_{\boldsymbol{w}} \ell)_{ij} = \frac{\partial (\nabla_{\boldsymbol{w}} \ell)_i}{\partial \boldsymbol{x}_j} \quad \text{and} \quad (\nabla_y \nabla_{\boldsymbol{w}} \ell)_i = \frac{\partial (\nabla_{\boldsymbol{w}} \ell)_i}{\partial y}. \quad (11)$$

3.3 Iteratively Reweighted FIL

Existing privacy mechanisms fail to provide equitable protection against privacy attacks for different subgroups [Yaghini *et al.*, 2019]. We address this issue with iteratively reweighted Fisher information loss (IRFIL; Algorithm 1), which yields a model with an equal per-example FIL across all examples in $\mathcal{D}$. This is done by re-weighting the per-example surrogate loss $\ell(\boldsymbol{w}^\top \boldsymbol{x}_i, y_i)$ over repeated computations of the minimizer $\boldsymbol{w}^*$. After the first iteration in Algorithm 1, the weight for the i-th example is inversely proportional to the initial η_i. At successive iterations, the per-example weight ω_i^t is multiplicatively updated by a value inversely proportional to the current model's FIL. The normalization on line 7 is primarily for numerical stability.

Model	$\bar{\eta}$	Accuracy
Linear	0.040 ± 0.014	100
+IRFIL	0.047 ± 0.000	99.8
Logistic	0.027 ± 0.012	99.8
+IRFIL	0.024 ± 0.000	99.7

Table 1: The mean $\bar{\eta}$ ($\pm$ standard deviation) of the example-level η and test accuracy before and after IRFIL.

4 Experiments

We compute the Fisher information loss for both linear and logistic regression models on MNIST. We perform binary classification of the digits 0 and 1 using a training dataset of 12,665 examples. We normalize all inputs to lie in the unit ball, $\max_i \|\boldsymbol{x}_i\|_2 \leq 1$, and then project each input using PCA onto the top twenty principal components for the corresponding dataset.

For linear regression we do not apply L_2 regularization, and for logistic regression, λ is set to the largest value such that the training set accuracy is the same up to two significant digits as that of linear regression. For linear regression, the targets are $y_i \in \{-1, 1\}$, and we compute the exact minimizer. For logistic regression, we use limited-memory BFGS to compute the minimizer. We typically report η assuming a Gaussian noise scale of $\sigma = 1$; hence $\eta = \|\boldsymbol{J}_f\|_2$. Code to reproduce our results is available at https://github.com/facebookresearch/fisher_information_loss.

Validating FIL. Histograms of the per-example η are shown in figure 2, separated by output class label. For both models, the histograms have distinct modes. The mode at the larger value is for the digit 0, implying that the model in general contains more information about images of 0 than of 1. Figure 3 shows the eight images with the largest and smallest η for the linear regression model. The images with the smallest η are consistent with the class means in table 1. These correspond to the digit 1 written in a very typical manner. As expected, the images with the largest η are much more idiosyncratic. Of the 100 largest η examples, 24 overlap between the logistic and linear regression models.

Reweighted FIL. We empirically evaluate the IRFIL algorithm in figure 4, which plots the standard deviation of the per-example η against the number of re-weighting iterations. The per-example η converge to the same value after only a few iterations for both linear and logistic regression. Table 1 shows the mean and standard deviation of η, as well as the test accuracy for models trained with and without IRFIL. Neither the average FIL $\bar{\eta}$ nor the test accuracy are especially sensitive to the IRFIL algorithm. However, without IRFIL the standard deviation in η is significantly higher, implying that the initial information leakage varies substantially across training examples. Overall, IRFIL achieves fairness in privacy loss with little change in accuracy or average privacy loss.

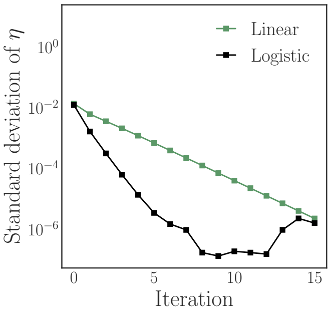


Figure 4: The standard deviation of the example-level η over iterations of the IRFIL algorithm.

5 Discussion and Future Work

We demonstrated that Fisher information loss can be used to assess the information leaked by a model about its training data. A primary benefit of FIL over *a priori* guarantees like differential privacy is the ability to measure information leakage at various granularities with respect to the data at hand. This also allows FIL to be used to construct models with equi-distributed leakage, which can be done with iteratively reweighted FIL. Furthermore, since FIL explicitly measures the inferential power of an adversary it can be used by practitioners to tailor the resulting privacy to the adversary's knowledge and capabilities.

We motivated the use of FIL via the Cramér-Rao bound and delineated the corresponding threat model. However, the assumption that the adversary is limited to unbiased estimators may not hold in the presence of auxiliary information. The implications of this should be further investigated. Furthermore, unlike differential privacy, FIL does not implicitly degrade with correlated data. This property of FIL should also be further studied.

The IRFIL algorithm closely resembles iteratively reweighted least squares (IRLS), which has been widely studied for ℓ_p-norm regression [Green, 1984] and sparse recovery [Daubechies *et al.*, 2010]. While IRLS often converges rapidly in practice, the theoretical convergence rates are difficult to derive and do not reflect the empirical results [Ene and Vladu, 2019]. We also observed rapid and robust convergence with the IRFIL algorithm without any hyper-parameter tuning. Future work may help understand the convergence of IRFIL from a theoretical standpoint.

Finally, we considered FIL in the common setting of output-perturbed generalized linear models with Gaussian noise. However, many possible extensions exist in the randomization used including alternative noise distributions such as the Laplace distribution [Dwork *et al.*, 2006], objective perturbation [Chaudhuri *et al.*, 2011], or quantifying leakage via predictions directly using, for example, the exponential mechanism [McSherry and Talwar, 2007]. Furthermore, extending FIL to the setting of non-linear and non-convex models will facilitate its utility and broader adoption.

Acknowledgments

Thanks to Mark Tygert for essential contributions to the ideas in this work and for providing copious feedback.

References

[Agrawal and Aggarwal, 2001] Dakshi Agrawal and Charu C Aggarwal. On the design and quantification of privacy preserving data mining algorithms. In *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 247–255, 2001.

[Anderson, 1977] Harald Anderson. Efficiency versus protection in a general randomized response model. *Scandinavian Journal of Statistics*, pages 11–19, 1977.

[Carlini *et al.*, 2019] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 267–284, 2019.

[Chaudhuri *et al.*, 2011] Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3), 2011.

[Cummings *et al.*, 2019] Rachel Cummings, Varun Gupta, Dhamma Kimpara, and Jamie Morgenstern. On the compatibility of privacy and fairness. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, pages 309–315, 2019.

[Daubechies *et al.*, 2010] Ingrid Daubechies, Ronald DeVore, Massimo Fornasier, and C Sinan Güntürk. Iteratively reweighted least squares minimization for sparse recovery. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 63(1):1–38, 2010.

[Dong *et al.*, 2019] Jinshuo Dong, Aaron Roth, and Weijie J Su. Gaussian differential privacy. *arXiv preprint arXiv:1905.02383*, 2019.

[Dwork and Roth, 2014] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[Dwork *et al.*, 2006] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

[Ekstrand *et al.*, 2018] Michael D Ekstrand, Rezvan Joshaghani, and Hoda Mehrpouyan. Privacy for all: Ensuring fair and equitable privacy protections. In *Conference on Fairness, Accountability and Transparency*, pages 35–47, 2018.

[Ene and Vladu, 2019] Alina Ene and Adrian Vladu. Improved convergence for ℓ_1 and ℓ_∞ regression via iteratively reweighted least squares. In *International Conference on Machine Learning*, pages 1794–1801. PMLR, 2019.

[Farokhi and Kaafar, 2020] Farhad Farokhi and Mohamed Ali Kaafar. Modelling and quantifying membership information leakage in machine learning. *arXiv preprint arXiv:2001.10648*, 2020.

[Farokhi and Sandberg, 2017] Farhad Farokhi and Henrik Sandberg. Fisher information as a measure of privacy: Preserving privacy of households with smart meters using batteries. *IEEE Transactions on Smart Grid*, 9(5):4726–4734, 2017.

[Fredrikson *et al.*, 2014] Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In *23rd USENIX Security Symposium (USENIX Security 14)*, pages 17–32, 2014.

[Ghosh and Kleinberg, 2016] Arpita Ghosh and Robert Kleinberg. Inferential privacy guarantees for differentially private mechanisms. *arXiv preprint arXiv:1603.01508*, 2016.

[Green, 1984] Peter J Green. Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *Journal of the Royal Statistical Society: Series B (Methodological)*, 46(2):149–170, 1984.

[Kay, 1993] Steven M Kay. *Fundamentals of statistical signal processing*. Prentice Hall PTR, 1993.

[Li *et al.*, 2007] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd International Conference on Data Engineering*, pages 106–115. IEEE, 2007.

[Liu *et al.*, 2016] Changchang Liu, Supriyo Chakraborty, and Prateek Mittal. Dependence makes you vulnberable: Differential privacy under dependent tuples. In *NDSS*, volume 16, pages 21–24, 2016.

[Long *et al.*, 2018] Yunhui Long, Vincent Bindschaedler, Lei Wang, Diyue Bu, Xiaofeng Wang, Haixu Tang, Carl A Gunter, and Kai Chen. Understanding membership inferences on well-generalized learning models. *arXiv preprint arXiv:1802.04889*, 2018.

[Machanavajjhala *et al.*, 2007] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):3–es, 2007.

[McSherry and Talwar, 2007] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103. IEEE, 2007.

[Mironov, 2017] Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 263–275. IEEE, 2017.

[Papernot *et al.*, 2017] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pages 506–519, 2017.

[Samarati and Sweeney, 1998] P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical report, 1998.

[Shokri *et al.*, 2017] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE, 2017.

[Tramèr *et al.*, 2016] Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction apis. In *25th USENIX Security Symposium (USENIX Security 16)*, pages 601–618, 2016.

[Yaghini *et al.*, 2019] Mohammad Yaghini, Bogdan Kulynych, and Carmela Troncoso. Disparate vulnerability: On the unfairness of privacy attacks against machine learning. *arXiv preprint arXiv:1906.00389*, 2019.

[Yeom *et al.*, 2018] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pages 268–282. IEEE, 2018.